Computing and Data in Climate Science

Kathy Yelick
Associate Dean for Research, Division of Computing, Data Science, and Society
Professor of Electrical Engineering and Computer Sciences
University of California, Berkeley

Senior Advisor on Computing, Lawrence Berkeley National Laboratory
Moore’s Law + Parallelism + $$
Moore’s Law

It’s hard to think exponentially

But it’s also hard to stop
Dennard Scaling is Dead; Moore’s Law Will Follow

Science implication: Atlas computing estimate off by $1B
Specialization: End Game for Moore’s Law

![Graph showing specialization from general purpose to special purpose with an increase in Ops/Joule]

- General Purpose
- GPUs
- Reconfigurable
- Special Purpose
Accelerators in the Top500
Data Movement is Expensive

Hierarchical energy costs.

- 6 pJ: Cost to move data 1 mm on-chip
- 100 pJ: Cost of a typical floating point operation
- 120 pJ: Cost to move data 20 mm on chip
- 250 pJ: Cost to move off-chip, but stay within the package (SMP)
- 2000 pJ: Cost to move data off chip into DRAM
- ~2500 pJ: Cost to move data off chip to a neighboring node

Image: http://slideplayer.com/slide/7541288/
Research for Climate Science

The global crisis needs cross-disciplinary teams

- Economics
- Computer Vision
- Natural Language
- RL & Control
- Earth Science
- Physics-Constrained Learning
- Sociology
- Causal inference
- Transfer learning
- Ecology
- Biology
- Time-series analysis
- Law and Public Policy
- Material Science
- Environmental Science
- Engineering
- Quantifying Uncertainty
- Ecology
- Economics
- Sociology
- Engineering
Faster Computers: More Detail

Michael Wehner, Prabhat, Chris Algieri, Fuyu Li, Bill Collins, Lawrence Berkeley National Laboratory; Kevin Reed, University of Michigan; Andrew Gettelman, Julio Bacmeister, Richard Neale, National Center for Atmospheric Research
Understanding Clouds

4D Stereophotogrammetry leads to new data sets, Rusen Oktem and David Romps

New mathematical models for simulation
Data Structures for Climate Modeling
Climate Domain Specific Languages

5-point Jacobi

Asymmetry near boundary

Red-black

4-color

Dawn: a High Level Domain-Specific Language Compiler Toolchain for Weather and Climate Applications
## Analytics vs. Simulation Kernels:

<table>
<thead>
<tr>
<th>7 Dwarfs of Simulation</th>
<th>7 Giants of Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Particle methods</td>
<td>Generalized N-Body</td>
</tr>
<tr>
<td>Unstructured meshes</td>
<td>Graph-theory</td>
</tr>
<tr>
<td>Dense Linear Algebra</td>
<td>Linear algebra</td>
</tr>
<tr>
<td>Sparse Linear Algebra</td>
<td>Sorting</td>
</tr>
<tr>
<td>Spectral methods</td>
<td>Hashing</td>
</tr>
<tr>
<td>Structured Meshes</td>
<td>Alignment</td>
</tr>
<tr>
<td>Monte Carlo methods</td>
<td>Basic Statistics</td>
</tr>
</tbody>
</table>

Phil Colella

NRC Report + our paper

Opportunities to Reduce Energy Use

Global energy consumption by sector

- Transportation 28%
- Industry 26%
- Residential 12%
- Commercial 10%
- Electrical System Losses 24%

Where are biggest impacts in reducing energy consumption?

Role of computing and data:

- Modeling engines, manufacturing processes, building materials
- Designing urban systems, transportation, and the power grid
- Use of reinforcement learning in optimizing these systems

Data from IEA based on 2019 data
Energy Efficiency in Industry

Paper industry is 4th Largest Energy Consumer in US

Chombo-Pulp: Apply adaptive embedded boundary solver to resolve flow around pulp fibers and in felt pore space

Adaptive mesh refinement and interface tracking
Reinforcement Learning for traffic

- 30% of U.S. energy use is in transportation
- Optimize for travel time, reduced fuel consumption, and improved air quality
- Smooth traffic flow is more energy efficient
- Adversarial multi-agent transfer learning used even with mixed autonomy traffic to smooth traffic

Alex Bayen, Civil and Environmental Engineering, EECS, UC Berkeley, Director of the Institute for Transportation Studies
Reinforcement Learning in Buildings

- Survey of 73 studies on RL in building energy systems
- Various papers control HVAC, hot water, windows, lighting and more

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-free</td>
<td></td>
</tr>
<tr>
<td>Policy Gradient</td>
<td>3 of 73</td>
</tr>
<tr>
<td>Value-Based</td>
<td>56 of 73</td>
</tr>
<tr>
<td>Actor-Critic</td>
<td>11 of 73</td>
</tr>
<tr>
<td>Model-based</td>
<td>3 of 73</td>
</tr>
</tbody>
</table>

Zhe Wang, Tianzhen Hong  Energy Technologies Area, LBNL  DOI: 10.1016/j.apenergy.2020.115036
Opportunities to Reduce Carbon in Production

Renewable sources still play a modest role

Role of computing and data

- Design of solar materials, wind turbines, hydrogen fuel cells
- Design and impact analysis of carbon capture and sequestration
- Understanding economic drivers

Data from IEA based on 2019 data
Materials Design for Renewables + Storage
Design of Materials for Batteries, Solar Panels and More

Kristin Persson, Material Science and Engineering, UC Berkeley and LBNL, Materials Project PI

- Use of Bayesian optimization for layered materials
- [Bassman et al, npj Computational Materials 2018]
Inverse Design with ML

Designing materials, proteins, and small molecules with ML

High-dimensional design using machine learning

Search for molecules using an autofocusing generative model: moves around the design space, guided by an oracle.

Clara Fannjiang and Jennifer Listgarten at NeurIPS ‘20
Importance of Energy Storage

- Grid-scale storage is critical for use of renewables (solar, wind, etc.)
- Better data collection and methods could inform policies and economics.
- Need to predict adoption rates and develop infrastructure of various technologies.

Technology readiness of grid-scale energy storage
Updated from Schmidt et al. (2017).

Dan Kammen, Energy Resource Group, UC Berkeley
Scrub Carbon with Metal Organic Frameworks

- Metal Organic Frameworks (MOFs) to capture carbon in natural gas plants.
- Uses steam to regenerate the MOF for repeated use, reducing energy required for carbon capture.
- Latest design removes >90% of CO₂ from flue gas and 6X more than current (amine) technology.
- Exploring MOF design space
  - Traditionally explore MOF design with expensive Density Functional Theory (DFT)
  - Accelerate exploration using ML (graph NNs, etc.) with Gonzalez group (EECS)

Jeff Long College of Chemistry / UC Berkeley and LBNL
Learning from graphical structure

Nicolas Swenson, Aditi S Krishnapriyan, Aydin Buluc, Dmitriy Morozov, Katherine Yelick
Parallelism in Graph Neural Nets

- GNN models are huge; sampling has large number of edges
- Treat as sparse linear algebra problem

<table>
<thead>
<tr>
<th>Name</th>
<th>Vertices</th>
<th>Edges</th>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>9.4M</td>
<td>231M</td>
<td>300</td>
<td>24</td>
</tr>
<tr>
<td>Reddit</td>
<td>232K</td>
<td>114M</td>
<td>300</td>
<td>41</td>
</tr>
<tr>
<td>Protein</td>
<td>8.7M</td>
<td>1.05B</td>
<td>128</td>
<td>256</td>
</tr>
</tbody>
</table>

Tripathy, Yelick, Buluc, Reducing Communication in Graph Neural Network Training, SC’20
Communication-Avoiding Matrix Multiply

- 2D algorithm: never chop k dim
- 3D: Assume + is associative; chop k, which is replication of C matrix

Matrix Multiplication code has a 3D iteration space
Each point in the space is a constant computation (*/+)

for i
  for j
    for k
      C[i,j] * A[i,k] + B[k,j]
**Cannibalization effect**: Increasing market penetration of solar and wind reduces their own unit revenues and value factors (VF).

- Wind market penetration reduces solar VF, but solar penetration increases wind VF.

David Zilberman, Department of Agricultural and Resource Economics, UC Berkeley
Over 57 million acres of grassland in California mostly used for ranching.

Organic addition can sequester 9 metric tons of CO$_2$ per acre per year.

May save 28 million tons of CO$_2$e annually using just 5% of California's rangelands.

Whendee Silver / CNR UC Berkeley
First-Time Science Analysis with MetaHipMer

- What happens to microbes after a wildfire? (1.5TB)
- What at the seasonal fluctuations in a wetland mangrove? (1.6 TB)
- What are the microbial dynamics of soil carbon cycling? (3.3 TB)
- How do microbes affect disease and growth of switchgrass for biofuels (4TB)
- Combine genomics with isotope tracing methods for improved functional understanding (8TB)

JGI-NERSC-KBase FICUS projects
KmerProf comparing metagenomes

1) K-mer Analysis
K-mer histogram

2) Distance metrics
Count-based: Jaccard Index
Abundance: Bray-Curtis

Migun Shakya  LANL and Steve Hofmeyr LBNL
Distributed Hashing / Histogramming

Repeat while more to read and/or exchange

$P_0$
- Input read partition
- Load & parse block
- Outgoing k-mers & metadata
- Local k-mer set
- Store
- Hash table partition

$P_1$

Done bulk-synchronously in MPI (all2allv)
Or asynchronously with remote put/get/RPC

$P_{N-1}$
- Input read partition
- Load & parse block
- Outgoing k-mers & metadata
- Local k-mer set
- Store
- Hash table partition

Marquita Ellis (alignment), Steve Hofmeyr (k-mer counting), et al
K-mer counting now in UPC++

- New version in UPC++ avoids barriers
- And it's simpler!

Steve Hofmeyr, Rob Egan, Evangelos Gerganas, leads on MetaHipMer software
K-mer Counting

K-mer counter on Summit. (Note scales -- red k-mer exchange time is roughly equal.)
- Over 100x speedup (including communication); results expected to be data- and machine-dependent

Israt Nisa, et al
Mitigation

- Energy Efficiency
- Renewable Energy
- Carbon Capture
- Economic Drivers

Adaptation

- Extreme Climate Events
- Resilient Infrastructure
- Economic Impacts
- Planning for Migration
Integrated models of climate and the environment combine features learned from data and known physical laws.
Data-driven models produce new insights into carbon cycling

- ML methods bridge the scales to quantify the effect of CO$_2$ on vegetation and ecosystem function
- E.g., Increase in biomass by 2100 shown based on increase in CO$_2$ levels

ML methods measure influence of soil moisture on photosynthesis.

Show previous models of photosynthesis activity based on satellite data are ~15% too high.

Trevor Keenan, Dept. of Environmental Science, Policy and Management / UC Berkeley and EESA / LBNL
Big Data, Big Model, and Big Iron

Deep learning results are smoother than heuristic labels
Achieved over 1 EF peak on OLCF Summit: Gordon Bell Prize in 2018

Thorsten Kurth, Sean Treichler, Joshua Romero, Mayur Mudigonda, Nathan Luehr, Everett Phillips, Ankur Mahesh, Michael Matheson, Jack Deslippe, Massimiliano Fatica, Prabhat, Michael Houston
Data Analytics via Supervised Learning

Classification

Classification + Localization

Object Detection

Instance Segmentation

Extending image-based methods to complex, 3D, scientific data sets is non-trivial!

Slide source Prabhat
Identifying Extreme Climate Events

Uses of machine learning to robustly identify extreme events without heuristics or thresholds for specific data sets

New statistical models to characterize extreme weather

Detect atmospheric rivers and quantifying uncertainty using ML and Bayesian statistics

Implementing a new jet stream detector in TECA (Toolkit for Extreme Climate Analysis)

- Risser et al. 2020
- Paciorek et al. in prep
- O’Brien et al. 2020

Loring, O’Brien & Elbashandy

Bill Collins, LBNL and Earth and Planetary Science, UC Berkeley, Cascade Project PI
Reduce Environmental Impact in Ag

Iterative random forests used for spatial interpolation needed for high resolution models

Multi-model data (left) from a farm in Arkansas (satellite, multispectral UAV, fertilizer, water, temperature, etc.)

Sensors data for regional precipitation (right) uses Sequential Imputation Algorithm for time-series data improves quality by including stations with incomplete data

James Ben Brown, Statistics, UC Berkeley and Biosciences, LBNL
ML for detailed ecosystem models

Use of Random Forest ML to determine role of water in ecosystem productivity

- Find early summer water is critical to ecosystem productivity throughout
- Specific impact dependent on vegetation type (grassland, deciduous, evergreen)

Haruko Wainwright, Nuclear Engineering, UC Berkeley and LBNL
Earth systems are nonstationary and nonlinear. How to predict the future?

And how to properly represent critical interactions and feedbacks in our models?
Hydrology: physics and data models

Physical models
- First principles, lumped or distributed

Learning through data
- Regression, support vector machine, NNs

Complex models with feedback, conservation laws, etc.

- Information theory for causal inference and delineation of critical time and spatial scales
- Sparse regression to “discover” governing equations from data
- Formulate empirical forecasts constrained by physics

Laurel Larsen, Geography and Civil and Environmental Engineering, UC Berkeley
Watershed decision support

Decision constrained by regulations, climate predictions, agriculture and urban demands, etc.

Reference: California Water Commission

- High fidelity physics models + observations are computationally expensive
- Using DL-based surrogates for in-the-field decisions
- LSTM-RNN for long term groundwater predictions

Julianne Mueller, Computational Research Division, LBNL
## Measuring Climate Change Impacts

<table>
<thead>
<tr>
<th>Sector</th>
<th>Estimates</th>
<th>Adaptation Addressed</th>
<th>Global Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Forestry</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Species loss</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Sea-level rise</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Energy</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Human amenity</td>
<td>Yes</td>
<td>~Yes</td>
<td>No</td>
</tr>
<tr>
<td>Morbidity and mortality</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Migration</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Crime and conflict</td>
<td>Yes</td>
<td>No</td>
<td>Maybe</td>
</tr>
<tr>
<td>Productivity</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Water consumption</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Pollution</td>
<td>Yes</td>
<td>Maybe</td>
<td>No</td>
</tr>
<tr>
<td>Storms</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

“Quantifying Economic Damages from Climate Change” Journal of Economic Perspectives, Fall 2018

Maximilian Auffhammer, International Sustainable Development, UC Berkeley

https://pubs.aeaweb.org/doi/pdf/10.1257/jep.32.4.33
Inequality and the Social Cost of Carbon

Assess the economic impact of climatic change on agriculture, health, energy use, etc.

- Basis for “zero-emission credits” (NY, IL)
- Electric utilities planning (CO, MN, WA)
- Policy analysis (Mexico and Canada)

Inequity impacts

- SCC increases ~2-3x when inequality over time is disentangled from inequality between regions
- Based on two known models

David Anthoff, Energy & Resources Group, UC Berkeley
Understand economic impacts of climate

- Help decision makers understand the economic impacts of climate change
- Productivity and income are negatively impacted by heat
- Poorest 60% of people in the world will bear the brunt of economics impacts

Sol Hsiang, Goldman School of Public Policy, UC Berkeley
SIML: Satellite Imagery with ML

- Remotely estimating socioeconomic and environmental conditions
- A single sharable encoding of satellite imagery
  - Generalizes across prediction tasks (e.g. forest cover, house price, road length)
  - Accuracy competitive with deep neural networks
  - Orders of magnitude lower computational cost
- Others need only fit a linear regression to their own ground truth data in order to achieve state-of-the-art SIML performance.

Data-Intensive Development

- Understand impacts and targeting microloans and other aid
  - Real-time measure of poverty based on cell phone data and satellite imagery
  - Changing labor markets, migration, conflict and violence
  - Welfare-aware ML: a framework for multi-objective optimization with noisy data, balancing social welfare maximization with traditional loss minimization

Josh Blumenstock, School of Information, UC Berkeley

ML paper by Esther Rolf, Max Simchowitz, Sarah Dea, Lydia T. Liu, Daniel Bjorkegren, Moritz Hardt, Joshua Blumenstock ArXiv 2020
Need for an Integrated ML Climate Platform

- Behavioral changes
- Technical solutions
- Economics constraints
- Physical laws
- Geopolitical factors
Three ingredients for machine learning

Data

Algorithms

Machines

Dense Connections.
Interactive Data Science for Earth

Jupyter meets the Earth

- Large-Scale Hydrologic Modeling
- CMIP6 climate data analysis: The World Climate Research Program’s Coupled Model Intercomparison Project
- Geophysical inversions

Runs in browser
Laptops to Supercomputers
Text
Code
Output

Part of the EarthCube NSF program

Fernando Pérez, Joseph Hamman, Laurel Larsen, Kevin Paul, Lindsey Heagy, Christopher Holdgraf, Yuvi Panda
Is deep learning the only application?

Cautionary tale from HPL
Integrated Facilities for Science

Interconnected facilities where data is acquired, stored, analyzed and served

- Computing and Data Facilities
- Embedded Sensors
- Experimental Facilities
- User Community

ESnet

Edge Computing for Science

- Sequencers
- Light Sources
- Telescopes
- Particle Detectors
- Microscopes
- Environmental Sensors
- Particle Detectors
- Environmental Sensors
Profound Impacts of Climate Change

“We are the first generation to feel the effect of climate change and the last generation who can do something about it.”
Barack Obama, Former US President
Extra Slides
Specialization, Yes            Accelerators, No!

- More cores
- More data parallelism
- Narrow data types
- More memory spaces
- CPUs in control
- CPUs communicate

Memory
Vision for the Future and Role of Data Science
ML-based data analysis, decision-making, control and design for a sustainable climate future for all

- Data-driven decision making encourages mitigation and smooths adaptation
- Data informs governments worldwide to anticipate major employment disruptions, migration, economics

- Data-informed policies encourage carbon farming / ranching
- Reduces wildfire risk, ocean impacts, and ensures fair water access with ML-optimized interventions

- ML controls factories to residences
- Manages the renewal-dominated grid

- ML-designed materials used in renewables, grid storage
- ML-designed materials capture carbon before emission

Economics and Policies
Managed Environment
Smart Grid
Green Energy Materials

Economics
Policies
Managed
Environment
Green
Energy
Materials

Smart
Grid
Communication Dominates: Dennard was too good

\[
\text{Time} = \# \text{flops} \times \gamma + \\
\# \text{message} \times \alpha + \\
\# \text{bytes comm} \times \beta + \\
\# \text{diff memory locs} \times \alpha_2 + \\
\# \text{memory words} \times \beta_2
\]

Data from Hennessy / Patterson, Graph from Demmel
Put Accelerators in Charge of Communication

Architecture and software are not yet structured for accelerated-initiated communication (Summit with NVLink between Power9 CPUs and NVIDIA GPUs)

Taylor Groves et al
Partnering with Policymakers

- Strong partners in California state government on climate
- Innovative governance models: e.g., Water Data Consortium
- A data driven policy approach
  - Open Data Portal: https://data.ca.gov
  - Other state entities: Air Resources Board, Environmental Health Hazard Assessment, California Natural Resources Agency
  - Governor’s Senior Advisor on Climate (UCB Alum)